28 April 2015

Accelerating molecular simulations of proteins using Bayesian inference on weak information

Alberto Perez[1‡], Justin L. MacCallum[2‡] and Ken A. Dill[1,3,4]

[1] Laufer Center for Physical and Quantitative Biology, Stony Brook University
[2] Department of Chemistry, University of Calgary
[3] Department of Chemistry, Stony Brook University
[4] Department of Physics & Astronomy, Stony Brook University

Correspondence to:

Alberto Perez (alberto@laufercenter.org), Ken Dill (dill@laufercenter.org)

[‡]Both authors contributed equally to this work.

Supplementary Materials:

Supplementary Methods

MELD + CPI overview

MELD (Modeling Employing Limited Data), which is a method for integrative structural biology(1), is the heart of our computational engine. Briefly, MELD incorporates semi-reliable experimental information (sparse, ambiguous and unreliable data) into a physics based (molecular dynamics) engine. In MELD, protein structures are inferred by combining the physical and experimental aspects using a Bayesian formulation:

$$p(x|D) = \frac{p(D|x)p(x)}{p(D)} \sim p(D|x)p(x) \qquad (1)$$

where x represents structures, D represents experimental data, p(x|D) is the probability of the structure given the data, p(D|x) is the likelihood of the data given the structure, p(x) is the probability distribution of structures from the atomistic force field model and p(D) is an irrelevant normalization factor here. Restraints are used to incorporate the data into simulations.

Our significant departure here is that we can harness much `weaker information' than the `strong information' that is usually available from experiments. Here, we use coarse physical insights (CPI) are used instead of experimental data. These insights come from our general knowledge of protein structure and they are chosen so that they have a parallelism with the data used in MELD: (1) CPIs are *uncertain* – we know much of the information they give are false positives, (2) the set of true positive CPIs is *sparse* – there is little amount of information guiding towards native structures and (3) CPIs are *ambiguous* – there are many conformations that could satisfy them.

MELD Workflow

Here is an overview of MELD; details are given elsewhere (1). MELD generates a Boltzmann-distributed ensemble from a physical force field and from external information:

$$E_{total} = E_{physics} + E_{information}. \quad (1)$$

MELD turns the external information into collective, multi-body restraints called collections, which provide a mechanism to deal with unreliable information. For each collection, the user specifies the active fraction, telling MELD how many informational restraints must be satisfied. The active fraction represents how much MELD should "trust" the external information. During each time step, MELD enforces the restraints that have the lowest energy and ignores the rest, driving the system towards the nearest basin compatible with that subset of restraints (See Fig. 1).

All restraints are always present in the simulations, but we only enforce the preset percentage corresponding to the lowest energy restraints for the current structure (for each collection). The set of restraints enforced is updated every time step. Note that for each conformation there is a unique set of restraints enforced: the ones resulting in the lowest restraints energy (since they are independent, this is achieved by calculating all restraint energies, sorting them and choosing the fraction with the lowest energy). But the relationship between springs to structures is not unique: large regions of conformational space can map on to the same set of enforced restraints. Thus, the restraints funnel whole regions of conformational space into narrower regions where restraints are satisfied. There are different such regions depending on what restraints are enforced (see Fig. 1). This procedure can be framed in Bayesian terms(1).

The H,T-REMD procedure is key to how the restraints affect the simulation. At high replica index we impose weak force constants to the restraints (vanishing at the highest replica index) and run simulations at high temperature. Under these conditions many restraints are unsatisfied (e.g. very stretched springs in the case of a distance restraint) with almost no energetic penalty, so the ensembles explored in these conditions are broad and it is very easy to go from a conformation that has one set of low energy springs to another conformation where the set of lowest energy springs is different (see Fig. 1). Conversely, at low replica index, restraint force constants are high and temperatures low. In these conditions most of the restraints are satisfied and the ensembles are tight and it will be difficult to sample conformations where the set of restraints enforced is different than for the current structure. Exchanging between different replicas is accomplished using a standard metropolis Monte Carlo scheme. Accordingly, only conformations that have a low global energy (low restraint energy and low force field energy) will make it down to the lowest replicas. Thus, many conformations where the force field is not compatible with the restraints will never be sampled at low temperatures.

In essence the ladder acts in a way that resembles a simulated annealing run in NMR-based structure determination. But, there are significant differences. In solution NMR, a large number of correct restraints pull the structure towards a unique basin. In MELD+CPI the number of enforced restraints is much more sparse and ambiguous (different enforced springs), translating into many different topologies that are

compatible with the restraints. Hence, simulating annealing would give many possible topologies and would be unable to identify the correct one. Our REMD ladder gives proper populations that obey detailed balance, which provides the principle for selecting the correct conformation.

Our method is freely available to download from github. It contains two parts and requires the freely available OpenMM package: https://github.com/maccallumlab/meld-openmm-plugin.git and https://github.com/maccallumlab/meld.git.

Populations from MELD are representative of the underlying force field

As a last step, we select structures by clustering populations. H,T- REMD produces Boltzmann ensembles for each replica. We cluster the lowest temperature replicas and select representative structures from the five most populous clusters (see SI Methods). The population of each cluster is directly related to the free energy; selecting by cluster population is equivalent to selecting by free energy.

The restraints serve to sculpt the energy surface providing localized funneling. Effectively, the simulations are sampling from the following potential function,

$$H = H_{\text{forcefield}} + H_{\text{Physical Insight}};$$
$$where:$$
$$H_{\text{Physical Insight}} = H_{\text{sse}} + H_{\text{hydrophobic}} + H_{strand\ pairing} + H_{\text{confinement}}$$

(4)

The restraints are defined so as to always contribute $H_{\text{Physical Insight}} \geq 0$. In regions of conformational space that satisfy the restraints, the restraint energy is zero and the potential function is only that of the forcefield itself. Given H, we can compute the relative populations of the regions that are compatible with our insights at the lowest temperature replica:

$$\frac{P_i}{P_j} = \frac{e^{-\beta E_i}}{e^{-\beta E_j}} \qquad (6)$$

On the one hand, imposing the restraints speeds up and focuses the sampling. On the other hand, for structures compatible with the restraints (i.e., that have zero restraint energy) our method of imposing the restraints does not perturb the ratio of populations that would be given by the force field alone. In the limit of converged sampling and perfect force field accuracy, if the original springs contain a subset compatible with the native state, running either MELD+CPI or unrestrained MD will yield the same lowest free energy structure, but MELD+CPI will find this structure much faster.

Source of restraints: MELD+COARSE PHYSICAL INSIGHTS (MELD+CPI)

For the MELD+CPI workflow the restraints are automatically generated based on the protein sequence and its predicted secondary structure (psipred (2), porter (3)). Different collections are specified according to general rules of thumb (see Kinds of Physical Insights below). Each collection has its own accuracy parameter (f_H) determining how many restraints each collection should be enforced at any time. Most of the restraints in these collections are expected to be wrong, but a fraction of them (at least f_H) are expected to be satisfied in the native state. The success of MELD+CPI is in the systematic derivation of restraints based on a sequence and a set of general insights.

Instead of enforcing all the restraints in the simulation at the same time, the MELD framework is able to produce structures that are both compatible with the force field and a fraction of the restraints. Example scripts are available from https://github.com/laufercenter/MeldExamples.git.

Types of physical insights used for structure prediction

1. Secondary structure predictions

We obtain secondary structure predictions from PsiPred (2) or Porter (3). We turn these secondary structure predictions into a set of geometric restraints (see (1)). Since we know from prior study that secondary structure predictions are typically about 80 percent accurate, we set our active-fraction criterion for the secondary structure restraints to 0.8—meaning that once 80 percent of the secondary structure restraints are satisfied, the rest are ignored.

These tools create multiple sequence alignments (MSA) based on the sequence and then use short residue windows inside a neural network to assign secondary structure preferences. In this way, they do not use structural homology.

2. Strand pairing

We add long-ranged restraints that drive the system to favor hydrogen bonds between β-strands. We add springs that enforce all possible hydrogen bonds between residues in server-predicted strands. The active fraction is set so that $0.65N_\beta$ restraints are satisfied, where N_β is the number of residues in predicted β-strands. The factor 0.65 comes from our prior statistical analysis of small globular proteins in the PDB. The restraints are enforced between N and O atoms in the pairing residues (see eq. 7).

$$E = \begin{cases} 0 & if\ (r \leq 3\text{Å}) \\ k(r-3)^2 & if\ (3 < r \leq 4\text{Å}) \\ k(2r-7) & if\ (4 < r) \end{cases} \quad (7)$$

Where k=250kJ/(mol•nm^2) and r is the distance between N—O atoms in the current structure. These kind of functional restraints are typical in MD packages(4). Here the heuristic imposes that if the candidate N and O atoms are closer than 3Å there is no restraint, beyond that and until 4Å the restraint energy increases quadratically and linearly beyond 4Å.

3. Hydrophobic contacts

Third, we add long-range restraints between all possible pairings of hydrophobic amino acids. We set the active fraction to $0.08N_{pairs}$, where N_{pairs} is the number of pairs of hydrophobic residues and the factor 0.08 comes from a prior statistical analysis of small globular proteins in the PDB. The way we introduce each restraint is as a flat bottom harmonic potential between the Cβ of the two residues (see eq. 8). Where the hydrophobic residues are Alanine, Valine, Leucine, Isoleucine, Phenylalanine, Tryptophan, Methionine and Proline. Note that the distances between Cβ pairs that we impose are large enough to allow side chain rearrangement without incurring a restraint penalty.

$$E = \begin{cases} 0 & if\ (r \leq 9\text{Å}) \\ k(r-9)^2 & if\ (9 < r \leq 11\text{Å}) \\ 2k(2r-20) & if\ (11 < r) \end{cases} \quad (8)$$

Where k=250kJ/(mol•nm^2) and r is the distance between a specific pair of Cβ atoms in the current structure. The heuristic imposes no restraint penalty if the selected pair of Cβ atoms is closer than 9Å (allowing plenty of freedom for side chains to reorient). The restraint energy increases quadratically beyond that until 11Å and linearly after that.

4. Loose enforcement of compactness

Folded proteins are compact. We enforce this loosely. In this term, we enforce that all residues should be inside a sphere with radius R chosen as follows:

$$R = (16.9 * ln(N) - 15.8)/2 \quad (9)$$

where R is the radius (in Å) of the confinement sphere and N is the number of residues in the protein. This compactness insight on its own is not very restrictive, but when combined with the hydrophobic and strand pairing it enhances high-contact-order interactions. We derived this functional form based on a set of small proteins.

5. Disulfide bridges

For three proteins, we tried a parallel experiment: one in which we enforce the native disulfide bonds and one in which they are not. But, first there is a force field issue: molecular mechanics simulations cannot spontaneously change between oxidized/reduced cysteines. This means that even bringing two residues that should be disulfide bridged together would be hard in the wrong state due to steric clashes, non bonded interactions. The presence of these disulfide bonds can be detected experimentally (e.g. by mass spectroscopy). We ran simulations on three proteins both in the reduced and oxidized states and have seen very significant differences, especially in population distributions. Hence, such restraints can significantly enhance predictions of native structures, when they are known.

6. Our physical insights have not been optimized

In this paper we have described some insights we derived for folding globular proteins, but they can be used for a variety of phenomena. They have not been completely optimized. However, what is important is that they reduce the size of the conformational space enough for the simulations to find the native state relatively efficiently. In principle, many more physical insights could we added. However, since the

evaluation of all the restraints has to be done at each timestep, at some point the method will become impossibly slow. Hence, it is important that the insights that are added have a signal/noise ratio that make them directive enough to compensate the computational overhead.

Simulation Details

Molecular Dynamics

We model the proteins in full atomistic detail, combined with the implicit-solvation model of Onufriev, Bashford, and Case (5). For the protein interactions, we used an in-house modified version (to be published separately) of the AMBER12SB force field (4) that adds a CMAP-like (6) correction to reproduce the balance between α and β regions of the Ramachandran plot (accessible through the git repository). All our simulations are 500 ns long (per-replica) unless otherwise noted. Initial conformations are fully extended as generated by the tleap (4) sequence command. We use the OpenMM suite of programs (7) with a 2 femtosecond (fs) time step and Langevin dynamics.

Replica Exchange Molecular Dynamics

For efficient conformational sampling, we use a Hamiltonian and temperature Replica Exchange Molecular Dynamics (H,T-REMD) sampling approach with 30 replicas. The temperature ranges from 300K in the lowest replica to 450K in the highest, increasing geometrically. The heuristic restraints weaken at higher temperatures. At low replica index force constants are strong (250 kJ/ mol/nm^2) and at high replica index, they are zero, changing exponentially from the lowest to highest replica.

Clustering into representative structures

At the ends of each simulation, we collect together the most similar structures into clusters, as is commonly done in structure predictions. We have used average-linkage clustering (8, 9) with an ε value of two, which is standard (10, 11). As input for the clustering, we took the five lowest-temperature replicas. The accuracy of clustering is tested by computing the RMSD of the centroid to the native state. To avoid situations of loops and termini disrupting the clusters, the clustering is done on the Cα carbons of residues having predicted secondary structures. For the comparison with the native state we consider the Cα of all residues excluding flexible termini, as is standard in the field. Table S1 contains a description of the residues used for each protein. We arbitrarily define a threshold in which structures closer to native than 4 Å are regarded as being within the native basin.

Effects of disulfide bridges

For three proteins, we ran simulations with and without disulfide bonds enforced. We were able to sample and identify the native state for only one of them in the absence of disulfide bonds (1ery, see table S4). Adding the disulfide-bridge information significantly reduces the size of the protein's conformational landscape, allowing us to identify native-like states in two cases (the third one is just slightly over our 4Å

threshold, see Table S4). Disulfide bridges can be routinely determined through mass spectroscopy(12).

In this work we enforced 80% of secondary structure predictions. When analyzing the last 250 ns of simulation at low temperatures, we found that we were satisfying 88% of the secondary structure predictions (averaging over the 20 proteins). The lowest individual average was 73% and the highest 96%. This results because secondary structures first need to nucleate before they propagate. For these small proteins, 80% of the sse probably included most of the nucleation points in the secondary structures, and the cost of extending the sse was not much. Hence, the overall percentage that is satisfied is higher than expected.

Indeed, when analyzing the agreement between predictions and native in our dataset we found that 90% (minimum in the protein set is 58%) of the native sse is present in the prediction, while 95% (minimum in the protein set is 78%) of the predicted sse are present in the native. This numbers are higher than what we expected and probably one of the reasons for our success while using 80%.

Supplementary Tables

PDBid	number of residues	residues for rmsd	protein name
1bdd	60	9 to 57	B domain of staphylococcal protein A
1dv0	47	3 to 42	C-terminal UBA domain of HHR23A
1ery	39	1 to 34	Pheromone ER-11
1fex	59	6 to 59	MYB-domain of human RAP1
1gh1	90	1 to 74	wheat nonspecific lipid transfer protein
1hp8	68	1 to 68	human P8-MTCP1,
1pou	71	1 to 71	OCT-1 POU-specific domain
1ubq	77	1 to 72	structure of ubiquitin
2a3d	73	1 to 73*	De novo designed three-helix bundle protein.
3gb1	56	1 to 56	B1 domain of streptococcal protein G
1mi0	65	9 to 65*	redesigned protein G variant NuG2
1fme	28	4 to 26*	structure of FSD-EY,
1lmb	92	6 to 85*	lambda repressor-operator complex
1prb	53	8 to 50*	Albumin-binding domain
2f21	38	10 to 32*	WW domain
2f4k	35	2 to 31*	Villin subdomain HP-35
2hba	52	1 to 52*	N-terminal Domain of Ribosomal Protein L9
2jof	20	3 to 18*	Trp-cage
2p6j	52	5 to 48*	Designed engrailed homeodomain variant UVF
2wxc	47	9 to 27, 35 to 46*	BBL

Table S1: Proteins used in this work. * denotes values taken from reference(13)

Protein	Residues RMSD	C1	C2	C3	C4	C5	FPT (ns)	BRMSD (Å)
2jof	1-20	**1.2**	4.2	5.1	4.2	3.0	0.2	0.6
2f4k	1-35	**1.5**	5.8	6.2	6.8	7.5	1	0.7
1dv0	3-42	**1.0**	6.3	4.3	3.8	4.8	1	0.9
1bdd	9-57	**2.5**	2.9	3.2	4.6	3.1	13	1.4
3gb1	1-56	7.9	3.6	**3.4**	6.7	9.7	10	1.4
1prb	7-53	**2.5**	8.7	3.8	8.8	10.9	3	1.4
2p6j	5-49	**2.7**	9.1	4.2	5.2	5.6	6	1.7
1mi0	8-65	**3.3**	5.8	4.5	5.1	5.3	29	1.8
2f21	1-38	**3.6**	5.0	9.9	11.0	6.7	25	1.8
1ery*	1-34	2.8	6.9	6.6	3.9	**2.1**	1	2.0
1gh1*	1-74	4.4	**2.7**	3.0	5.4	3.9	6	2.0
1fme	1-28	7.7	7.5	7.1	4.5	**3.4**	0.4	2.0
2a3d	1-73	**2.9**	4.9	5.9	6.2	6.0	8	2.1
1ubq	1-72	**4.0**	6.6	5.4	5.3	7.6	58	3.0
1fex	6-59	8.6	4.9	6.3	**3.5**	8.8	21	3.2
2hba	1-52	**7.8**	9.8	9.8	8.2	10.2	61	0.9
2wxc	8-47	9.7	5.5	9.2	7.6	**5.4**	28	2.2
1pou	1-71	**8.9**	10.7	8.9	13.2	10.3	205	2.9
1hp8*	1-68	4.5	5.5	4.9	**4.3**	6.3	57	3.2
1lmb	8-92	**9.2**	12.2	9.9	12.1	10.8	57	3.6

Table S2: RMSD table between identified clusters and native. Best free energy RMSD in bold (best top 5 cluster: C1-5). The structures are sorted according to best ensemble rmsd (BRMSD). A line separates the structures that have not been identified through clustering. Here, only residues in flexible termini are excluded from the RMSD calculation. FPT: first passage time. * refers to proteins for which disulfide bridges were also included.

SSE	HC (8%)	SP	C1	C2	C3	C4	C5	Best	FPT (ns)
yes	no	no	11.9	8.7	10.0	10.9	10.7	3.2	447
no	yes	yes	10.8	11.6	12.1	11.0	11.0	8.2	-
yes	yes	no	11.2	13.4	10.8	10.0	10.8	4.9	-
yes	no	yes	6.3	11.7	10.2	11.2	6.8	3.7	366
yes	yes	yes	4.0	6.6	5.4	5.3	7.6	3.0	58
yes	Yes, 18%	yes	8.8	4.8	9.3	10.7	11.1	2.4	129

Table S3. Effect of restraints on Ubiquitin folding. SSE: secondary structure heuristic, HC: hydrophobic contact heuristic, SP: strand pairing heuristic, C1-C5: top 1 to 5 clusters rmsd; Best: rmsd of the best sampled structure, First Passage time (FPT): first time structure under 4Å is detected in the simulations.

PDB	with disulfides		without disulfides	
	Top5-RMSD	Best-RMSD	Top5-RMSD	Best-RMSD
1ery	2.1	2.0	3.8	2.1
1gh1	2.7	2.0	9.3	5.5
1hp8	4.3	3.2	7.3	3.6

Table S4: Including correct disulfide bond information improves structures.

pdb id	Number of residues	C1	C2	C3	C4	C5	Best-RMSD
1i6z	136	17.7	11.1	8.5	15.0	10.0	6.0
1lpe	145	22.2	16.5	8.9	20.4	12.7	8.5
1lre	82	10.6	10.6	11.1	8.3	9.6	3.6

Table S5: Results for non-globular proteins. HC: hydrophobic contact heuristic, SP: strand pairing heuristic, C1-C5: top 1 to 5 clusters RMSD; BRMSD: RMSD of the best-sampled structure.

28 April 2015

Supplementary Figures

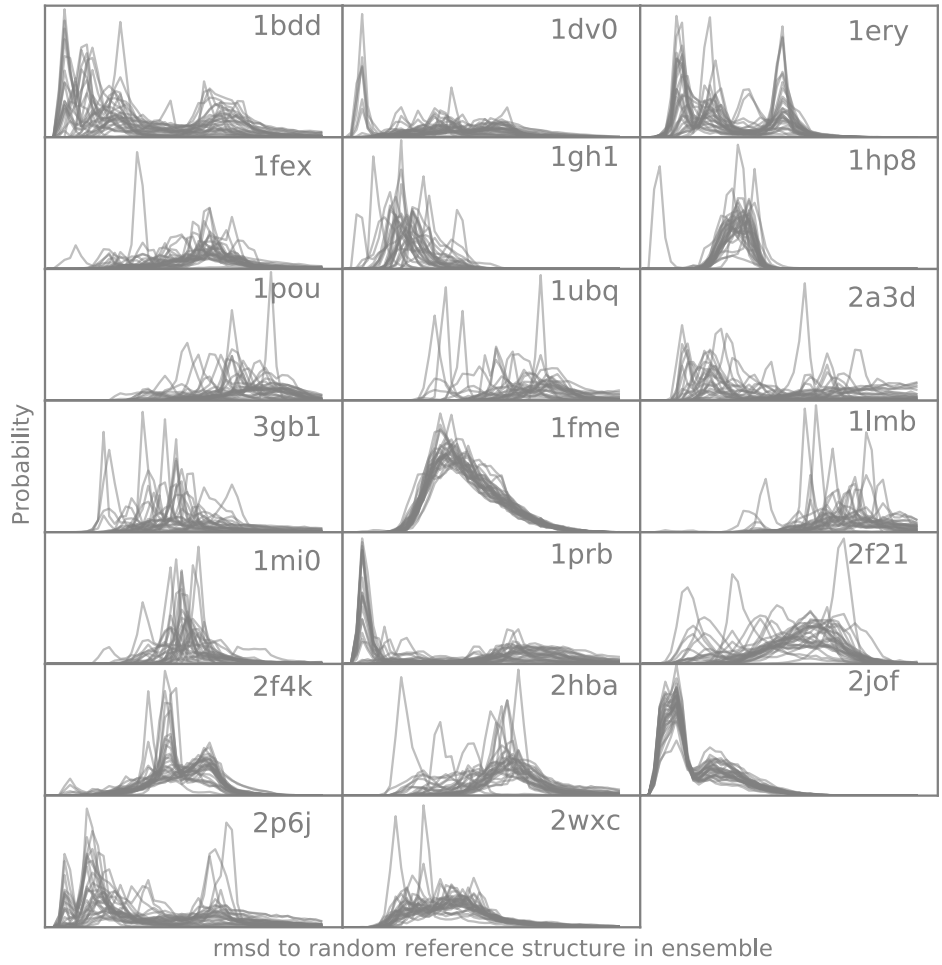

Figure S1: Not all simulations are converged. Each panel is for one of our 20 targets. The lines show 30 RMSD histograms, one for each "walker" in the REMD simulation (covering all replica conditions). The RMSDs are relative to the last frame of the simulation (this quality analysis does not require native structures). Converged simulations would give overlapping histograms. Some simulations (e.g. 1fme, 2jof) appear to be nearly converged, while others (e.g. 1hp8, 1pou) are not.

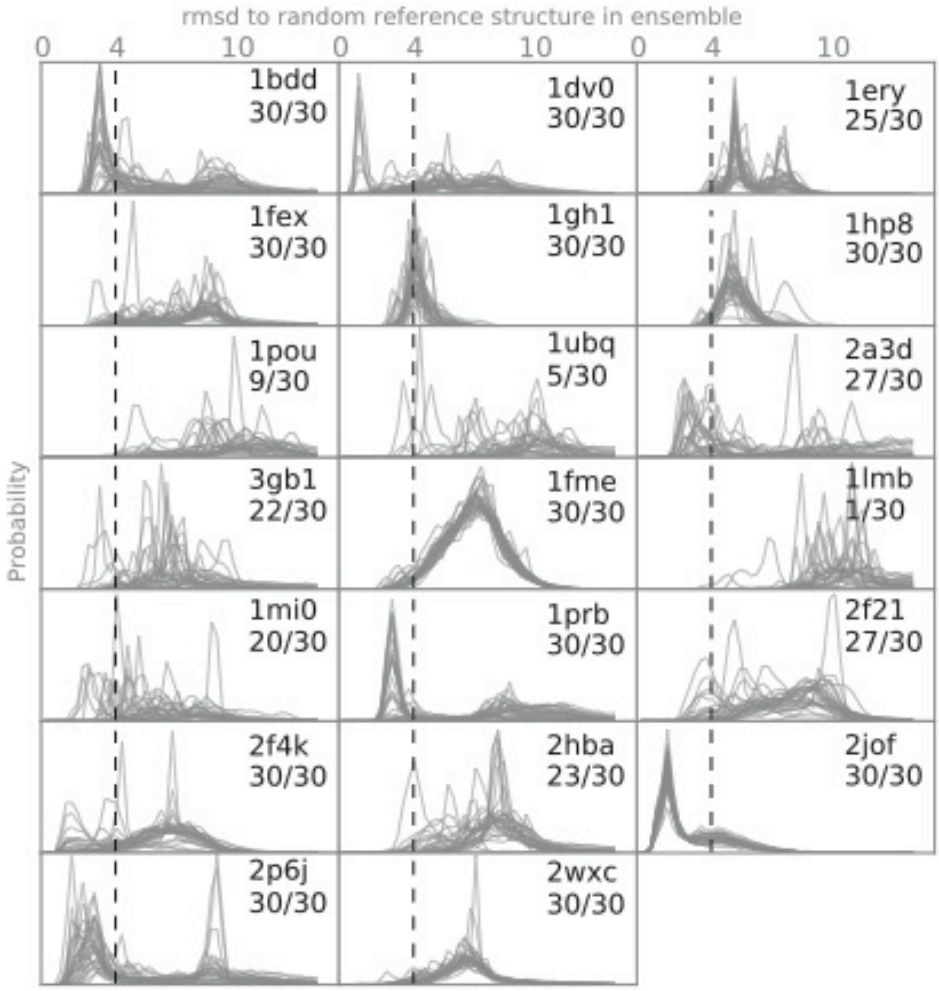

Figure S2: Multiple independent folding trajectories. Each panel is for one of the 20 target proteins. The lines show 30 RMSD histograms, one for each "walker" in the REMD simulation as they go up and down the replica exchange ladder. The RMSDs are relative to native (hence this analysis cannot be used to check for convergence in the way that figure S5 can). The graphic shows the number of replicas (X) that have folded (RMSD<4Å) out of the 30 possible (X / Y).

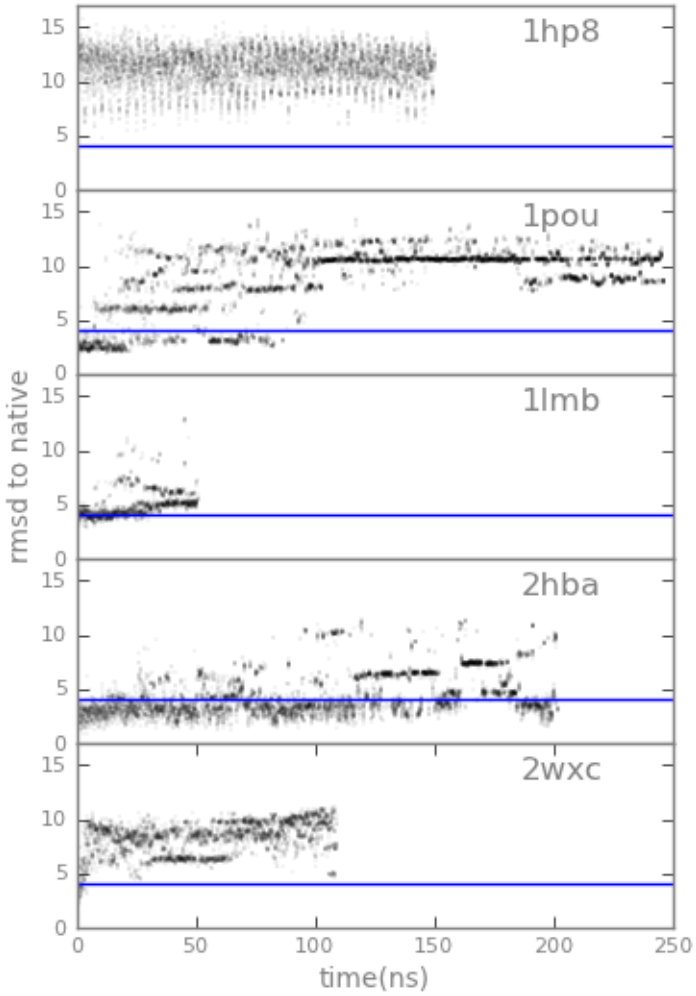

Figure S3: Unstable proteins with current force field. RMSD of the lowest temperature replica in H,T-REMD starting from native for five targets that we could not identify through clustering. The figure shows four of these proteins unfolding and one remaining stable (2hba). The blue line denotes the 4Å criteria for native basins.

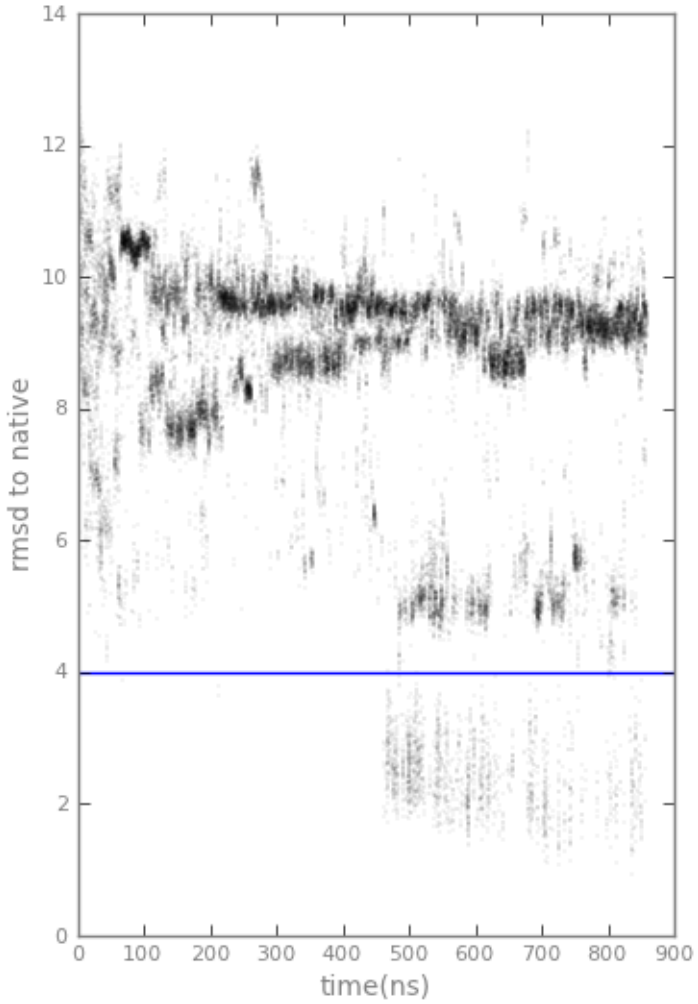

Figure S4. 2HBA requires longer simulation time. RMSD of the lowest temperature replica in an 850 ns MELD+CPI simulation of 2hba starting from extended. Population of native like structures significantly goes up after ~450ns, hinting at a convergence issue. The blue line denotes the 4Å criteria for native basins.

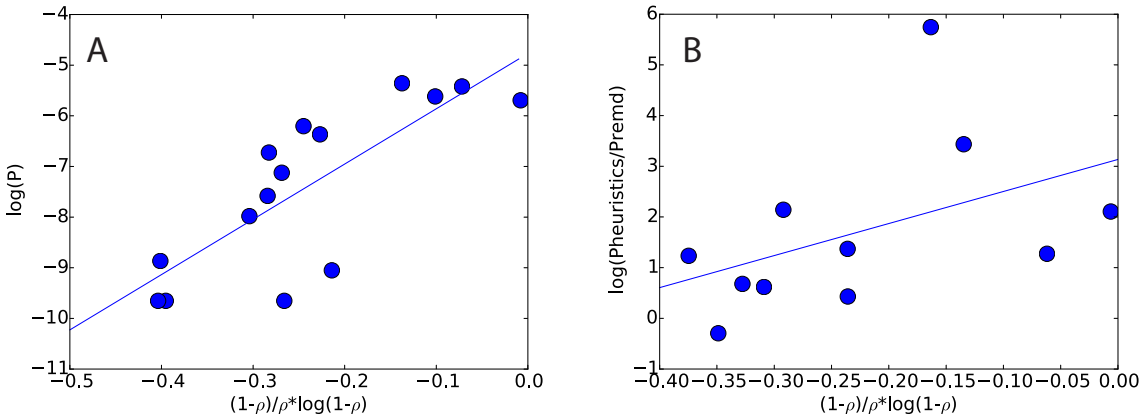

Figure S5. Flory-Huggins theory. A: For three proteins (protein G, 2HBA and ubiquitin) simulations were performed with different amounts of heuristic restraints. The plot shows the increase in performance with the normalized ratio of the number of springs and protein size. **B:** Comparison of increased performance with number of springs as explained with the Flory-Huggins theory. The dataset are the proteins that overlap between the current study and that of reference(13).

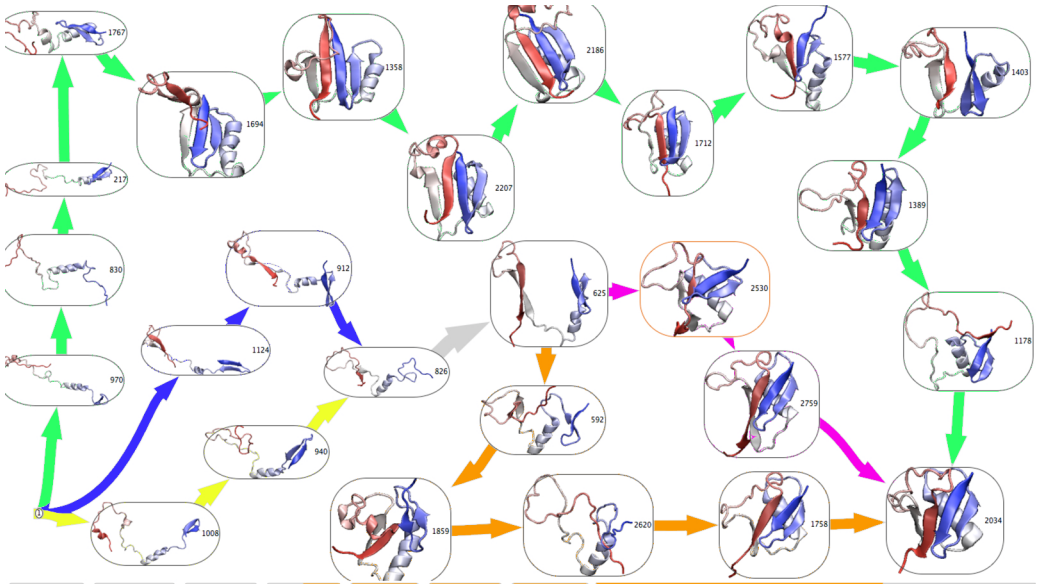

Figure S6. Representative MELD+CPI pathways in ubiquitin folding. On the bottom left we start from a fully extended conformation and in the bottom right we have the folded state. These are the highest flux pathways going from extended into the native state.

REFERENCES

1. MacCallum JL, Perez A, Dill K (2015) Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc Natl Acad Sci USA* 112(22):6985–6990.

2. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*.

3. Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 21(8):1719–1720.

4. Case DA, al E (2012) *Amber12* (University of California San Francisco).

5. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* 55(2):383–394.

6. MacKerell AD, Feig M, Brooks CL (2004) Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25(11):1400–1415.

7. Eastman P, et al. (2013) OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J Chem Theory Comput* 9(1):461–469.

8. Roe DR, Cheatham TE III (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 9(7):3084–3095.

9. Shao J, Tanner SW, Thompson N, Cheatham TE (2007) Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J Chem Theory Comput* 3(6):2312–2334.

10. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How Fast-Folding Proteins Fold. *Science* 334(6055):517–520.

11. Daura X, Gademann K, Jaun B (1999) Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition* 38(1/2):236–240.

12. Wu J, Watson JT (1997) A novel methodology for assignment of disulfide bond pairings in proteins. *Protein Science* 6(2):391–398.

13. Nguyen H, Maier J, Huang H, Perrone V, Simmerling C (2014) Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J Am Chem Soc* 136(40):13959–13962.